Cloud based Big Data Platforms and Tools for Data Analytics in the Big Data Engineering Curriculum

Yuri Demchenko, EDISON Project
University of Amsterdam

ICDATA'19
30 July 2019, Las Vegas

EDISON
building the data science profession

FAIRSFAIR
Fostering Fair Data Practices in Europe

mates

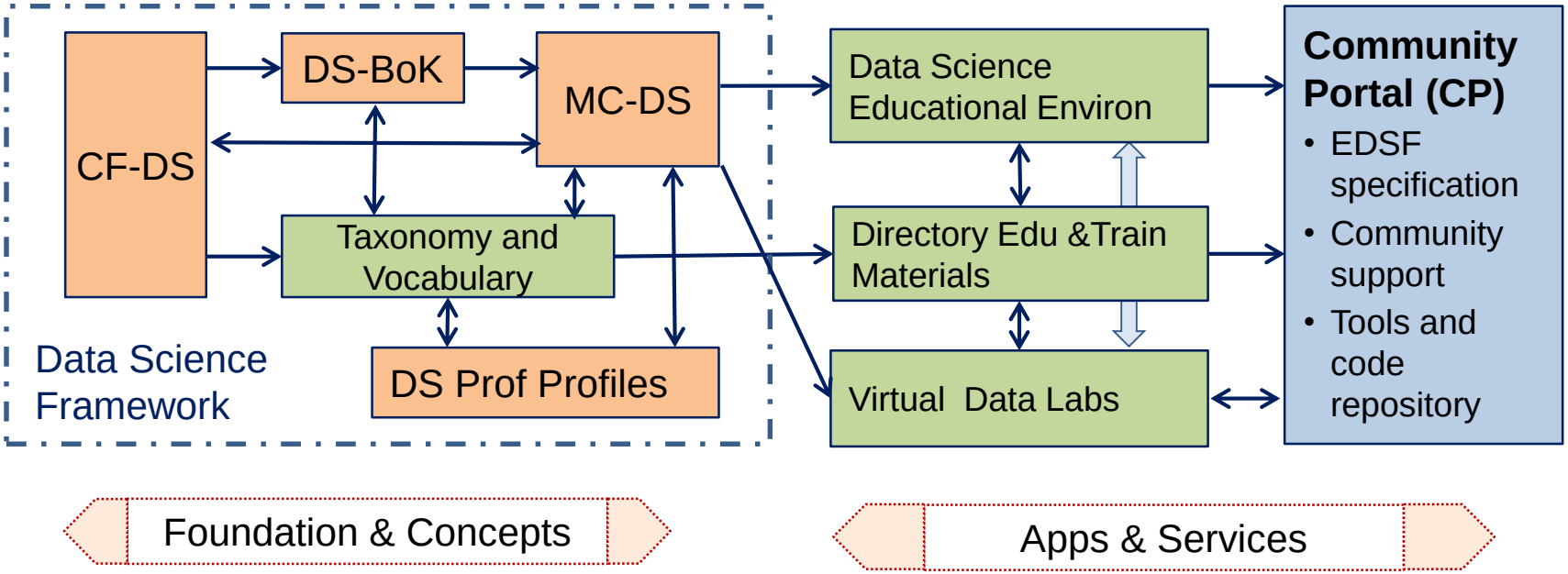
Outline

- Introduction
  - EDISON Data Science Framework (EDSF) and Data Science Engineering Body of Knowledge
  - BDIT4DA Syllabus example
- Big Data Infrastructure Technologies and Providers overview
  - Big Data Service Providers: AWS, Microsoft Azure, Google Cloud Platform
- Hadoop ecosystem and basics
  - Hadoop core components, YARN, Tez, LLAP
  - Applications for Big Data processing: Hive, Pig Latin, HBase
- NoSQL and modern cloud based SQL databases: Overview and practice
- Practice and Project: Working with Hadoop cluster
  - Recommended platforms: AWS, Azure, Cloudera Starter
  - Labs: Cloud, Cloudera Hue, HDFS, Hive, Pig
  - Group project
- Summary and further development

BDIT4DA Development and Implementation

- BDIT4DA is an implementations of the Data Science Engineering BoK and Model Curriculum
  - Developed as a part of EDISON Data Science Framework
- Using in real teaching at few universities
  - University of Amsterdam
  - University of Stavanger and joint course with Purdue University
  - Institute for Product Leadership (IPL), India
  - National Technical University of Ukraine “Igor Sikorski Polytechnic Institute”
- Partly implemented in the online programme of the Laureate Online Education (University of Liverpool)

EDISON Data Science Framework (EDSF) – Core components and community maintained services



EDISON Framework core components and documents

- CF-DS – Data Science Competence Framework (Part 1)
- DS-BoK – Data Science Body of Knowledge (Part 2)
- MC-DS – Data Science Model Curriculum (Part 3)
- DSPP – Data Science Professional profiles (Part 4)
- Data Science Taxonomies and Scientific Disciplines Classification

Applications and Services

- Virtual Data Science Labs
- Data Science Educational Environment
- Directory of edu & train resources
- Community Portal – currently github

Data Science Body of Knowledge (DS-BoK)

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- **Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)**
- **Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)**
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science Engineering Knowledge Area Group KAG02-DSENG

- KA02.01 (DSENG/BDIT) Big Data infrastructure and technologies, including NOSQL databased, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approached to big data analysis and common MapReduce algorithms;
- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with focus on data warehouse and data centers.

BDIT4DA course: Format and activities

- 8 lecture and practice sessions of 3 hours (mixed format – remote and face-to-face)
  - Interactive discussions
- Hands-on exercises as homework and during face-to-face sessions
- Practical assignments: AWS experience and data processing on Hadoop platform
- Group project to design company's Big Data Infrastructure and Data Analytics workflow
- Final course report including project report and fulfilled practical assignment

- Self-study part
  - Watch on-line tutorials
  - Read additional literature on the lecture topics
  - Experiment with tools

Example BDIT4DA Course Syllabus: Lectures/Modules/Sessions

Lecture 1 Cloud Computing foundation and economics.

Cloud service models, cloud resources, cloud services operation, multitenancy. Virtual cloud datacenter and outsourcing enterprise IT infrastructure to cloud. Cloud use cases and scenarios for enterprise. Cloud economics and pricing model.

Lecture 2 Big Data architecture framework, cloud based Big Data services

Big Data Architecture and services. Overview major cloud based Big Data platform: AWS, Microsoft Azure, Google Cloud Platform (GCP). MapReduce scalable computation model. Overview Hadoop ecosystem and components.

Lecture 3 Hadoop platform for Big Data analytics

Hadoop ecosystem components: HDFS, HBase, MapReduce, YARN, Pig, Hive, Kafka, others.

Lecture 4 SQL and NoSQL Databases

SQL basics and popular RDBMS. Overview NoSQL databases types. Column based databases and their use (e.g. HBase). Modern large scale databases AWS Aurora, Azure CosmosDB, Google Spanner.

Lecture 5 Data Streams and Streaming Analytics

Data streams and stream analytics. Spark architecture and components. Popular Spark platforms, DataBricks. Spark programming and tools, SparkML library for Machine Learning.

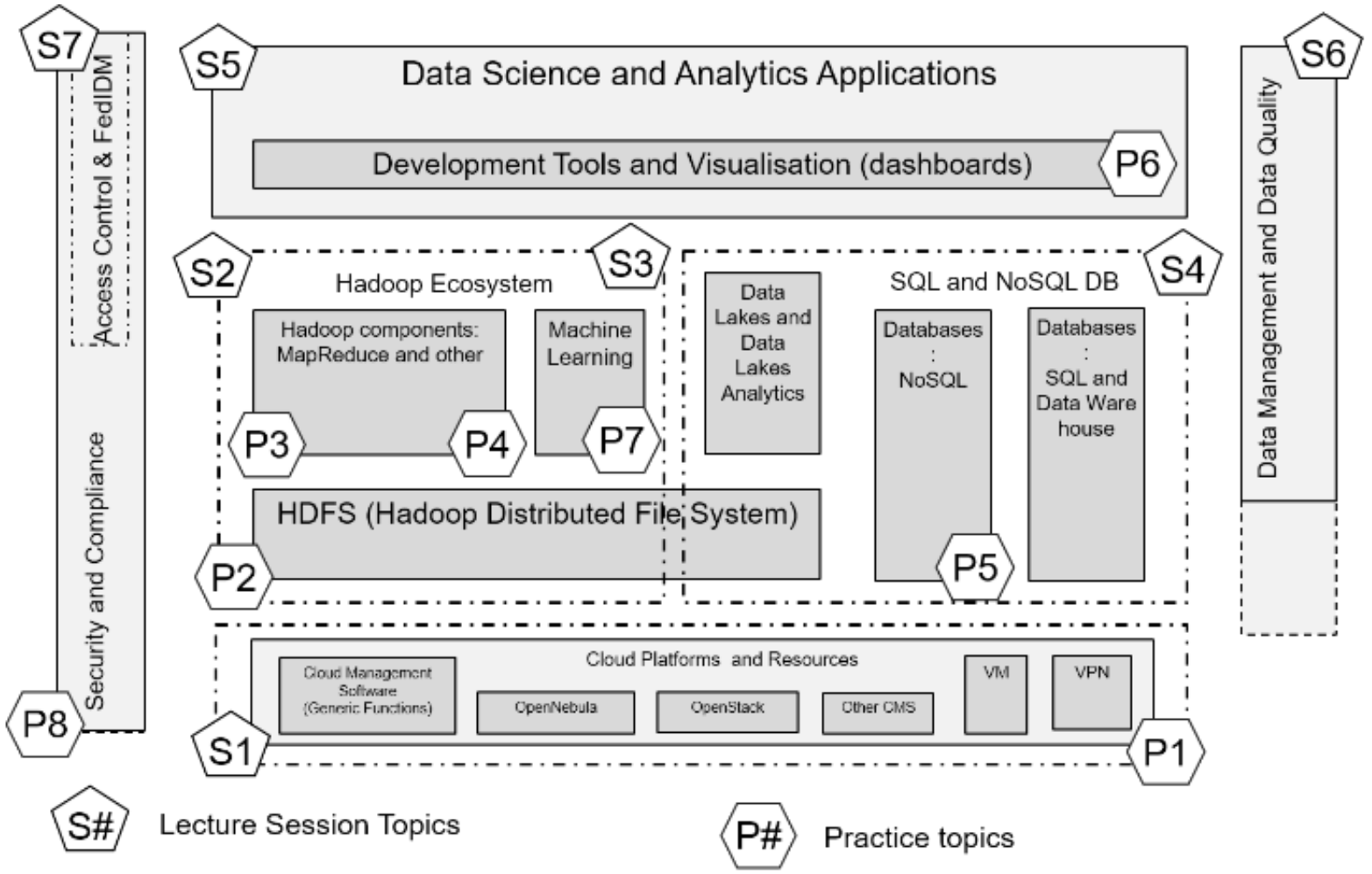
Lecture 6 Data Management and Governance.

Enterprise Big Data Architecture and large scale data management. Data Governance and Data Management. FAIR Principles in data management.

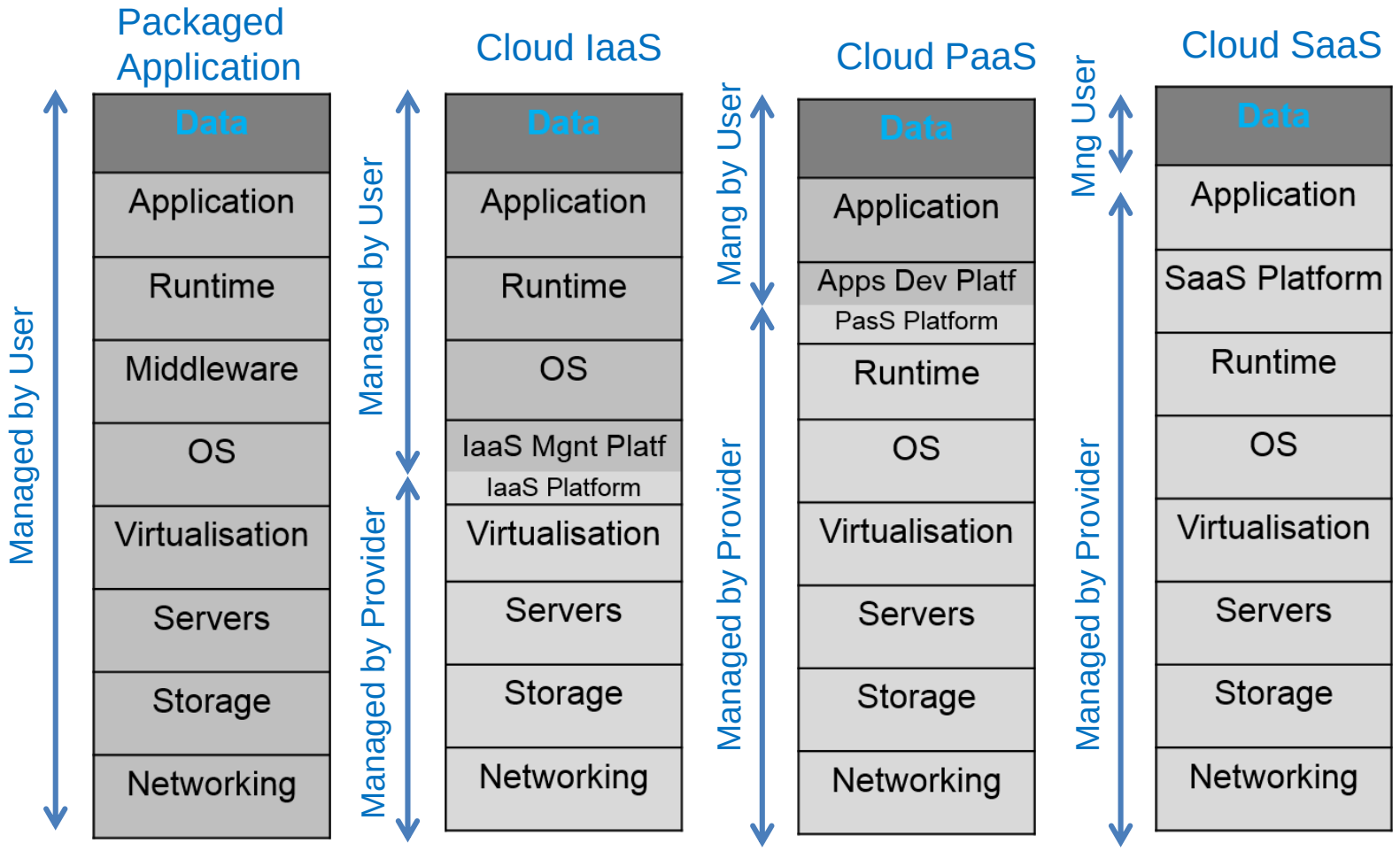
Lecture 7 Big Data Security and Compliance.

Big Data Security challenges, Data protection. cloud security models. Cloud compliance standards and cloud provider services assessment. CSA Consensus Assessment Initiative Questionnaire (CAIQ) and PCI DSS cloud security compliance.

Big Data Infrastructure Components and Course Modules



Data Security: Shared responsibility models when processing Data on Cloud



Note: Data always remain under user responsibility, however it may be processed on clouds. Physically, they may be processed at each of S/P/IaaS level

Module 2: Big Data Architecture Framework and Big Data platforms

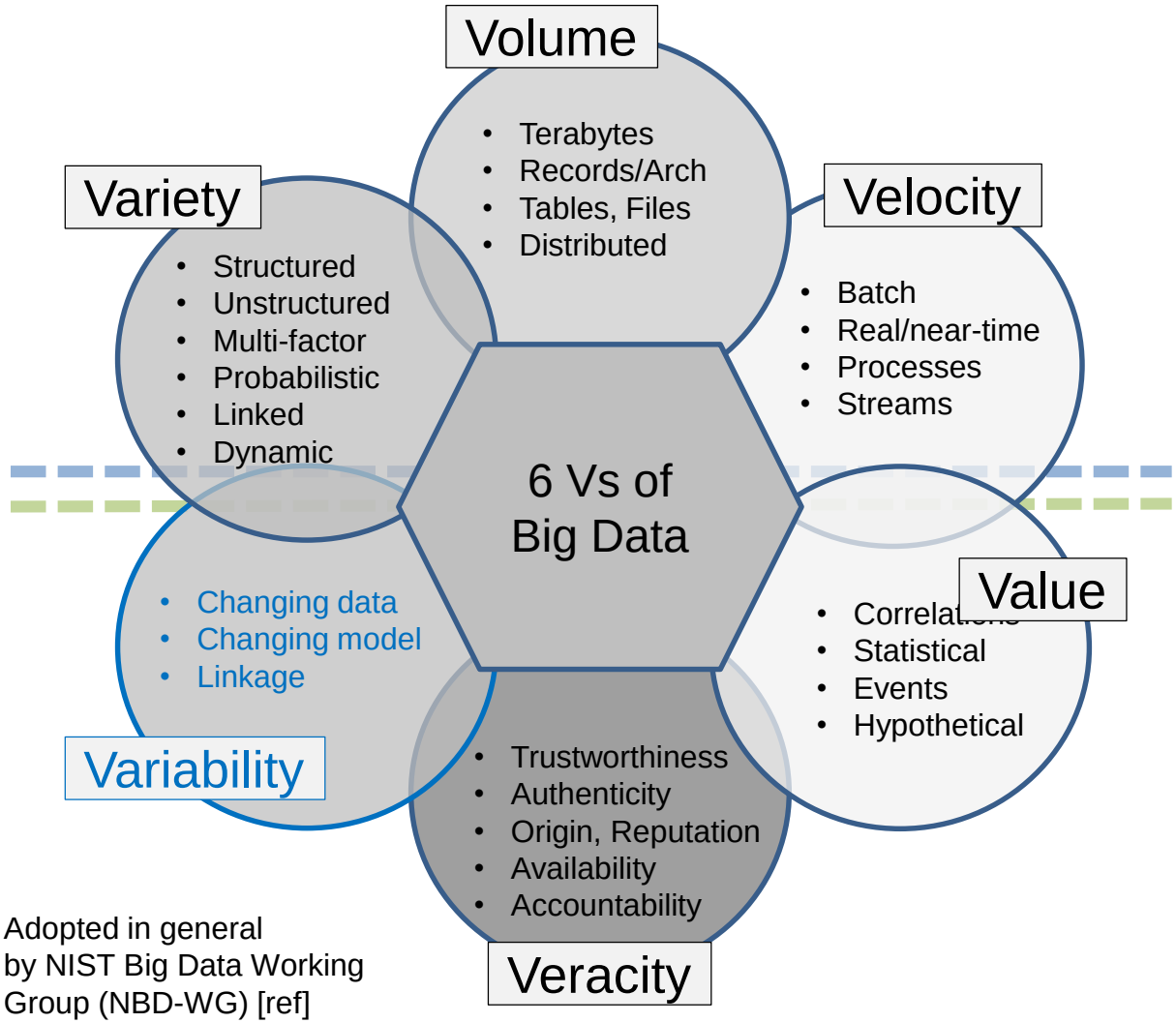
- Big Data definition and cloud based Big Data Infrastructure
  - Big Data Reference Architecture
- Big Data use cases
- MapReduce computation model
- Apache Hadoop Ecosystem
- Cloud based storage for Big Data
- Big Data Platforms and Providers
  - AWS Big Data services
  - Google Cloud Platform (GCP) Big Data services
  - Microsoft Azure Analytics Platform and HDInsight

Big Data and multiple sources of data

- Social Media
- IoT
- Internet
- Science
- Industrial data
- Communication, voice

Data analytics blending with open and social media data

Big Data Properties: 6 (3+3) V's of Big Data



Adopted in general by NIST Big Data Working Group (NBD-WG) [ref]

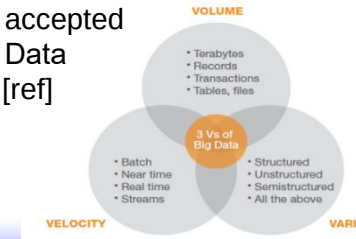
Generic Big Data Properties

- Volume
- Variety
- Velocity

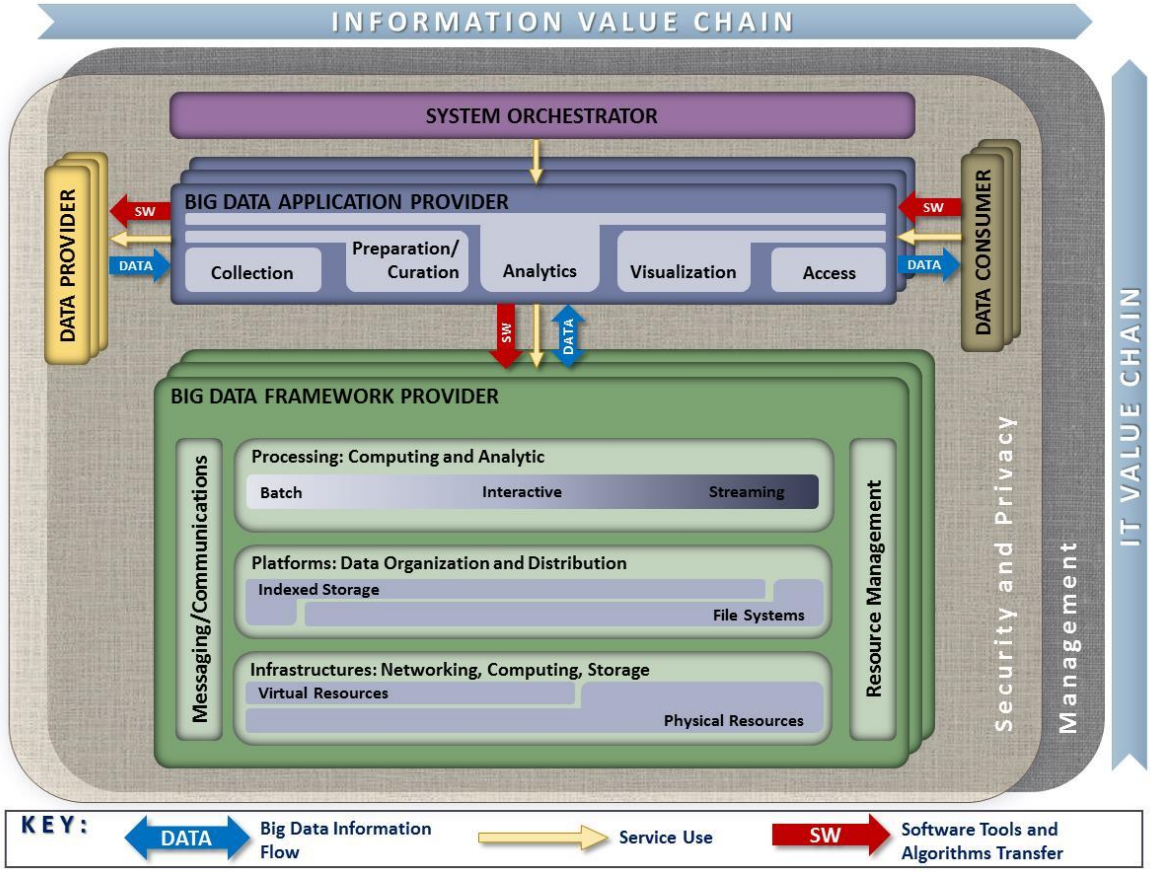
Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data by Gartner [ref]



NIST Big Data Reference Architecture (2018)



Main components of the Big Data ecosystem

- Data Provider
- Big Data Applications Provider
- Big Data Framework Provider
- Data Consumer
- Service Orchestrator

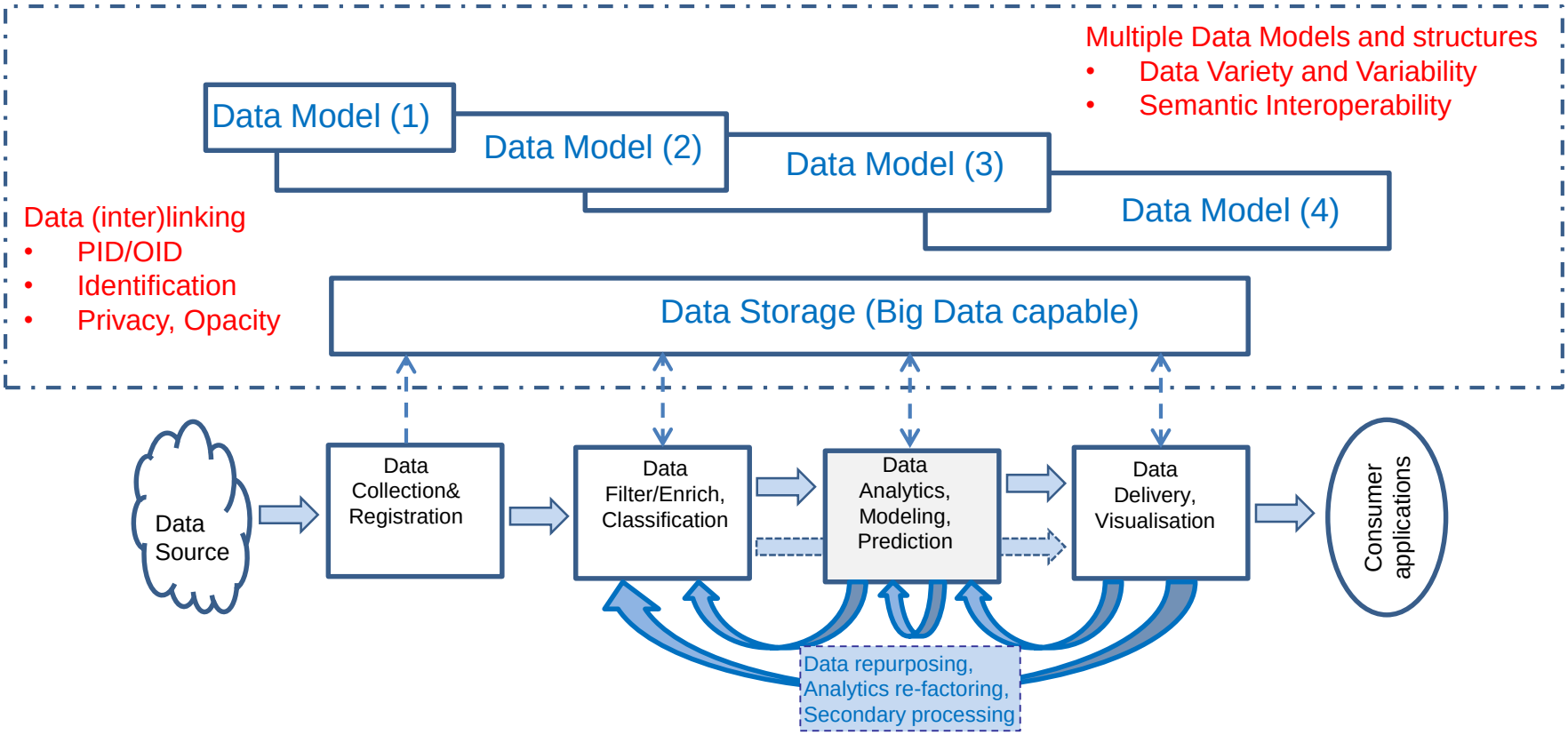
Big Data Lifecycle and Applications Provider activities

- Collection
- Preparation
- Analysis and Analytics
- Visualization
- Access

Big Data Ecosystem includes all components that are involved into Big Data production, processing, delivery, and consuming

[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

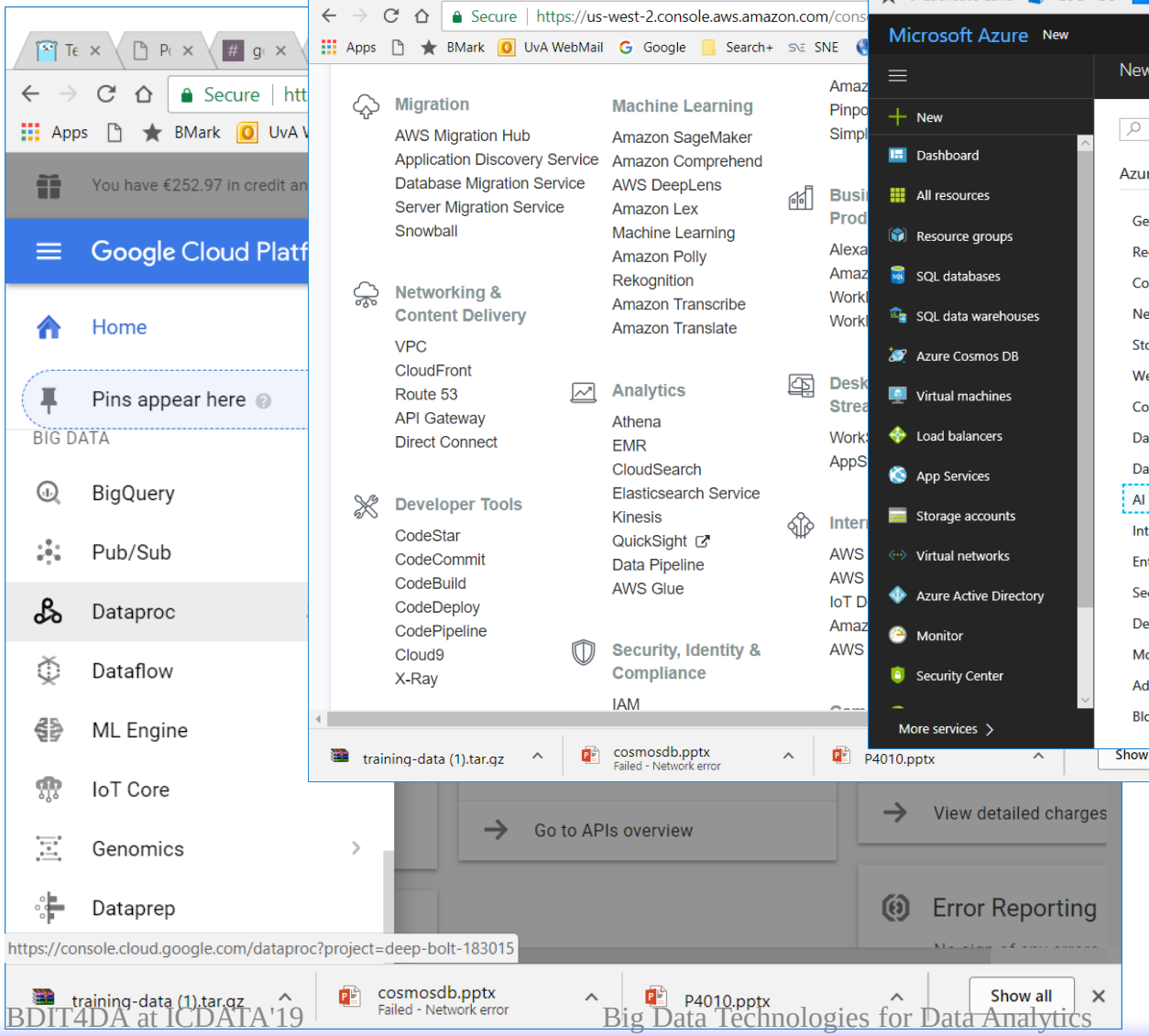
Data Lifecycle/Transformation Model



- Data Model changes along data lifecycle or evolution
- Data provenance is a discipline to track all data transformations along lifecycle

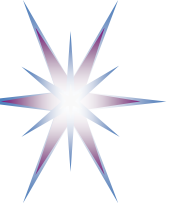
- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity

Big Data Stacks: Google ML, AWS EMR, Azure HDInsight



Cloud Resources: AWS Educate and Starter accounts

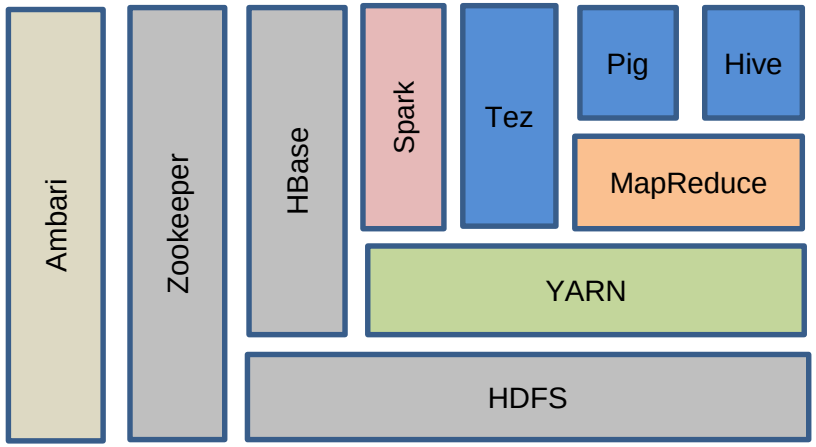
- AWS Educate:
  - Educational accounts are provided on request by teacher with 50USD credits per student
- AWS
  - 100 USD free credits for 1 year
  - Free tier services for 1 year
- Azure
  - 200 USD for 1 month
  - Free tier services for 1 year
- Google Cloud Platform
  - 300 USD for 1 year
  - Generous free tier services for 1 year



Module 3: Hadoop ecosystem: Important technology aspects

- Apache Hadoop Ecosystem
- HDFS
- MapReduce and YARN
- Tez, Zookeeper
- Hive: Architecture and Programming Model
- Pig Latin: Architecture and Programming Model
- Spark and Streaming Analytics

Apache Hadoop (Release 2.2+)
Current Releases 3.1 and 3.2 February 2019



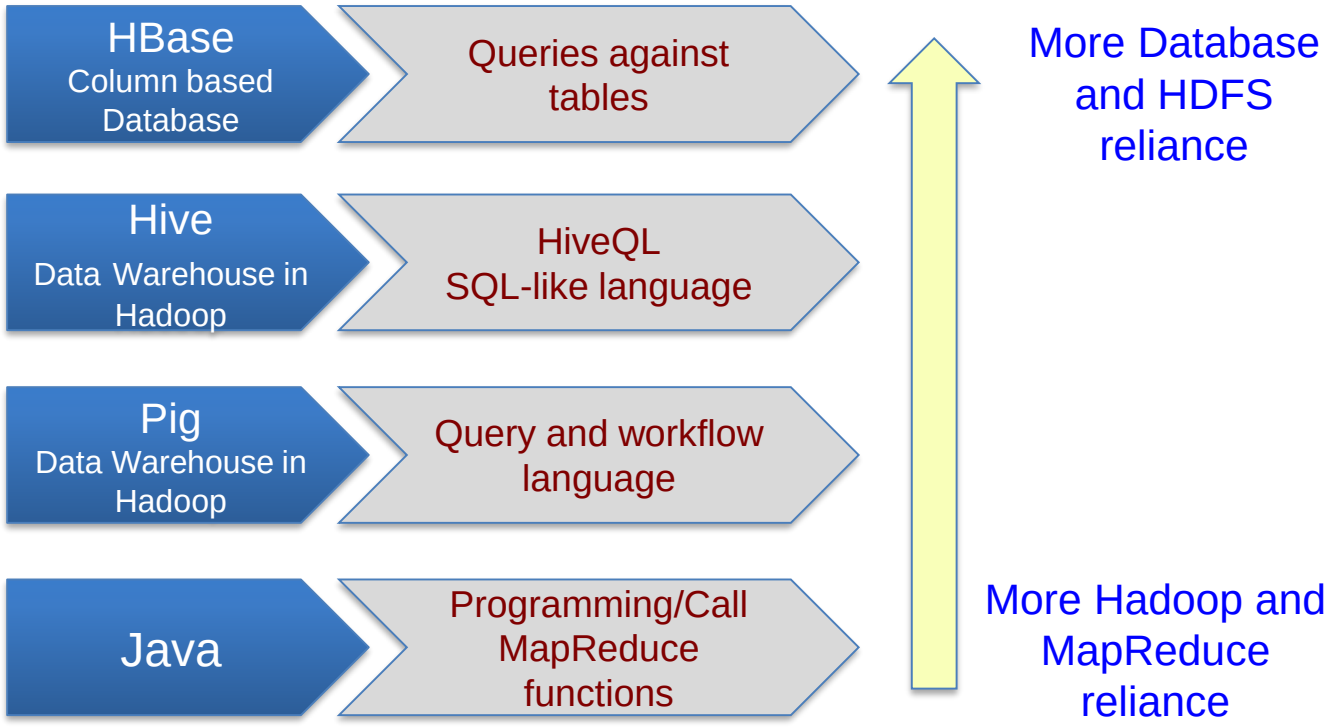
Apache Hadoop software stack includes the following main modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules and includes utilities and drivers to support different computer cluster and language platforms.
- **HDFS**: Hadoop Distributed File System optimized for large scale storage and processing of data on commodity hardware
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Hive:** A data warehouse system that provides data aggregation and querying.
- **Pig:** A high-level data-flow language and execution framework for parallel computation.
- **HBase:** A distributed column oriented database that supports structured data storage for large tables
- **Tez:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- **ZooKeeper:** A scalable coordination service for distributed applications.
- **Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Ambari:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters
- **Cassandra:** A scalable multi-master database protected against hardware failure
- **Mahout:** A scalable machine learning and data mining library.
- **Avro:** A data serialization system that supports rich data structures

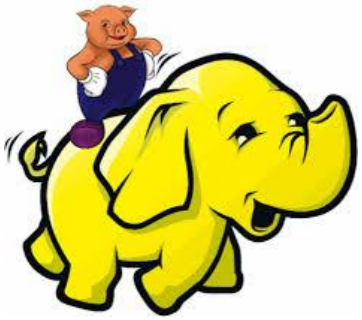
Module 4: Query Languages for Hadoop



- **Java:** Hadoop's Native Language
- **Pig:** Query and Workflow Language
- **Hive:** SQL-Based Language
- **HBase:** Column-oriented Database for MapReduce

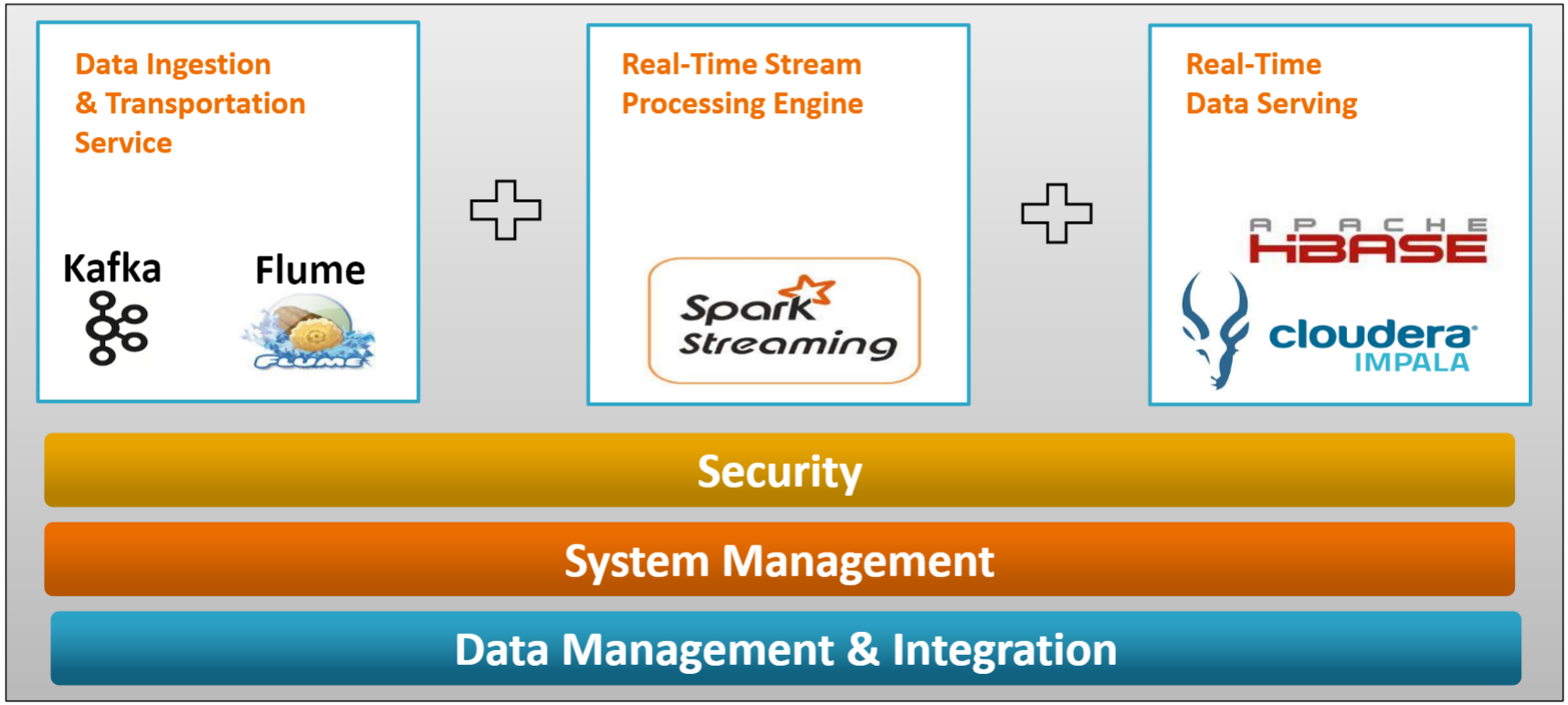
Hive, Pig Latin, Oozie

- Solution: Provide higher-level data processing languages
- Hive: Data warehousing application in Hadoop
  - Query language is HQL, variant of SQL
  - Tables stored on HDFS as flat files
  - Developed by Facebook, now open source
- Pig: Large-scale data processing system
  - Scripts are written in Pig Latin, a dataflow language
  - Developed by Yahoo!, now open source, Roughly 1/3 of all Yahoo! internal jobs
- Oozie
  - Server-based workflow scheduling system to manage Hadoop jobs.
  - Workflows defined as a collection of control flow and action nodes in a directed acyclic graph.
- Common idea:
  - Provide higher-level language to facilitate large-data processing
  - Higher-level language “compiles down” to Hadoop jobs



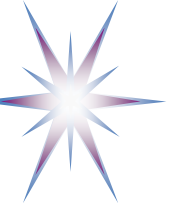
Spark and Streaming Analytics

Key Components of Streaming Architectures



cloudera

 5



Module 5: NoSQL Databases and New Cloud based SQL Databases

- Data types and data models
- SQL databases: ETL and ELT processes
- Distributed systems: CAP theorem, ACID and BASE properties
- NoSQL databases overview
- Modern cloud databases and CAP challenges

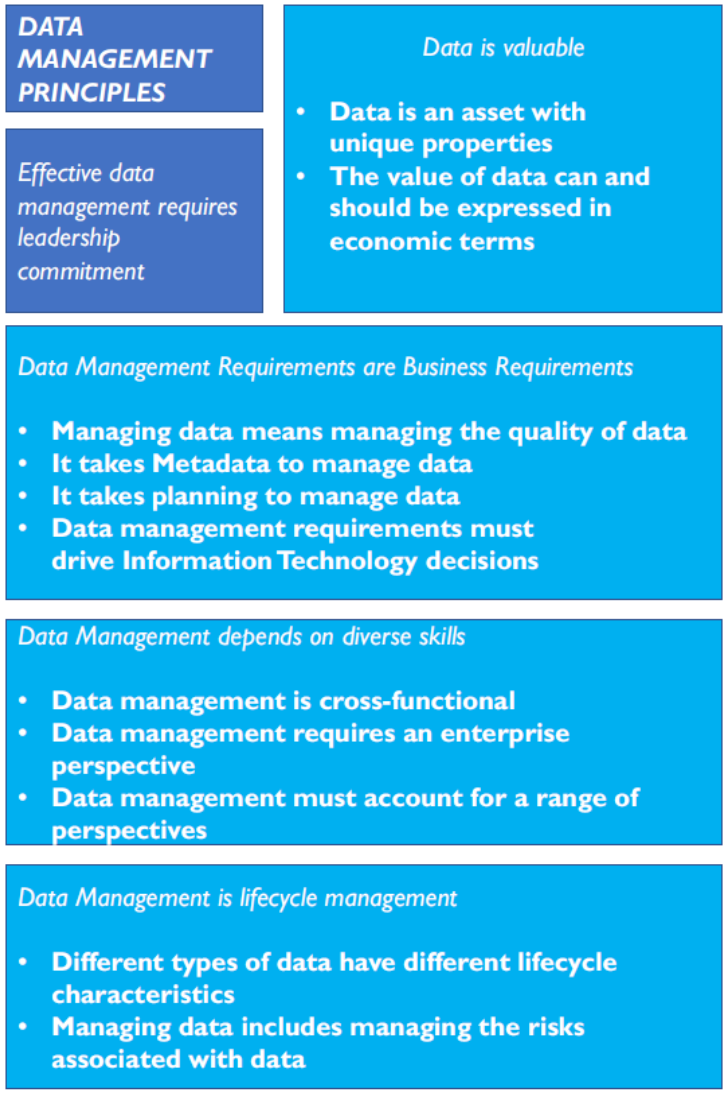
Module 6: Data Management and Governance

- Based on DMBOK by DAMA, NIST BDRF, CIMI Data Management Maturity model

Knowledge Areas describe the scope and context of data management activities.

1. **Data Governance** provides direction and oversight for data
2. **Data Architecture** defines the blueprint for managing data assets by aligning with organizational strategy to establish strategic data requirements and designs to meet these requirements.
3. **Data Modeling and Design** is the process of discovering, analyzing, representing, and communicating data requirements in a precise form called the *data model*.
4. **Data Storage and Operations**
5. **Data Security**.
6. **Data Integration and Interoperability**
7. **Document and Content Management**
8. **Reference and Master Data**
9. **Data Warehousing and Business Intelligence**
10. **Metadata**
11. **Data Quality**

DMBOK Data Management Principles



- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- Effective data management requires leadership commitment

BDIT4DA Practice

Practice 1: Getting started with the selected cloud platform. For example, Amazon Web Services cloud; cloud services overview EC2, S3, VM instance deployment and access.

Practice 2: Understanding MapReduce, Pregel, other massive data processing algorithms. Wordcount example using MapReduce algorithm (run manually and with Java MapReduce library).

Practice 3. Getting started with the selected Hadoop platform. Command line and visual graphical interface (e.g. Hue), uploading, downloading data. Running simple Java MapReduce tasks.

Practice 4. Working with Pig: using simple Pig Latin scripts and tasks. Develop Pig script for programming Big Data workflows. This can be also done as a part of practical assignment on Pig.

Practice 5. Working with Hive: Run simple Hive script for querying Hive data base. Import external SQL database into Hive. Develop Hive script for processing large datasets. This can be also a part of practical assignment on Hive.

Practice 6: Streaming data processing with Spark, Kafka, Storm. Using simple task to program Spark jobs and using Kafka message processing The option for this practice can also use Databricks platforms that provides a good tutorial website.

Practice 7: Creating dashboard and data visualisation. Using tools available from the selected Hadoop platform to visualise data, in particular using results from Practice 5 or 6 that is dealing with large datasets where dashboard is necessary

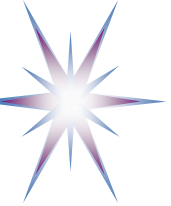
Practice 8. Cloud compliance practicum. This practice is important for the students to understand the complex compliance issues for applications run on cloud. Using Consensus Assessment Initiative Questionnaire (CAIQ) tools.

Group Project: Enterprise Big Data Infrastructure

- Developed by group of 3-5 students
- Must address all aspects of BDI for a hypothetical SME/SMB company
  - Business model and data processing flow
  - BDI components and Data Analytics platform and tools
  - Data Management Plan (DMP) and organisational roles
  - Security and Compliance
- Project report

Assessment

- Primarily is project based
  - May include exam – then less requirements for the project
- Grade structure
  - Project based: Project 70%, Practice 30%
  - Project + Exam: Exam 60%, Project 30%, Practice 10%
- Literature study is reported as part of the project report



Ongoing development

- Spark labs with Hadoop (EMR or HDInsight)
- Visualisation and dashboard with Hadoop
- Full practice set with the Azure HDInsight

Questions and Discussion

- Open to share experience
- Open to cooperation

- Continuing development to catch new technologies and trends

EDISON Initiative Online Presence

- EDSF github project - https://github.com/EDISONcommunity/EDSF
  - Component documents CF-DS, DS-BoK, MC-DS, DSPP

- EDISON Community work area and discussions - https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
- Mailing list - edison-net@list.uva.nl

- EDISON project website - old domain ***edison-project.eu*** expired: Legacy information to be moved to http://edison-project.net/

Links to EDISON Resources

- EDISON Data Science Framework Release 3 (EDSF)
  https://github.com/EDISONcommunity/EDSF

Component EDSF documents

CF-DS – Data Science Competence Framework
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_CF-DS-release3-v09.pdf

DS-BoK – Data Science Body of Knowledge
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DS-BoK-release3-v04.pdf

MC-DS – Data Science Model Curriculum
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_MC-DS-release3-v04.pdf

DSPP – Data Science Professional profiles
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DSPP-release3-v05.pdf